

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

Google scholar

duplicate document webcrawler OR crawler rep

Search

[Advanced Scholar Search](#)[Scholar Preferences](#)[Scholar Help](#)
Scholar [All articles](#) - [Recent articles](#) Results **1 - 10** of about **1,110** for **duplicate document webcrawler OR crawler representative OR proxy** (0.16 seconds)

Design and implementation of a distributed **crawler** and filtering processor

D Zeinalipour-Yazti, M Dikaiakos - Lecture notes in computer science, 2002 - Springer

... executed repeatedly until all links of the **document** at hand ... added to the URL-Queue, dropping all **duplicate** URL's ... that have been visited by the **crawler** already ...

[Cited by 21](#) - [Related articles](#) - [Web Search](#) - [BL Direct](#) - [All 7 versions](#)

[PDF] •On the evolution of clusters of near-**duplicate** web pages

D Fetterly, M Manasse, M Najork - Proceedings of the 1st Latin American Web Congress, 2003 - cwr.cl

... that have been found to be near-**duplicates** of one ... the data using the Mercator web **crawler** [12], customized ... by whitespace, and then segmented the **document** into 5 ...

[Cited by 62](#) - [Related articles](#) - [View as HTML](#) - [Web Search](#) - [All 21 versions](#)

Engineering a multi-purpose test collection for web retrieval experiments- *psu.edu (pdf)

P Bailey, N Craswell, D Hawking - Information Processing and Management, 2003 - Elsevier

... is their inability to detect "near-**duplicates**", ie pages ... or whether they were missed by the **crawler** for some ... since we applied a fixed **document** cutoff of ...

[Cited by 120](#) - [Related articles](#) - [Web Search](#) - [All 10 versions](#)

The Case of the **Duplicate** Documents Measurement, Search, and Science- *unimelb.edu.au (pdf)

J Zobel, Y Bernstein - Lecture Notes in Computer Science, 2006 - Springer

... helpful to have multiple copies of a **document** in an ... the presence of **duplicates** can indicate a **crawler** failure. ... a qualitative definition of what a **duplicate** was ...

[Cited by 6](#) - [Related articles](#) - [Web Search](#) - [BL Direct](#) - [All 9 versions](#)

Managing duplicates in a web archive- *ul.pt (pdf)

D Gomes, AL Santos, MJ Silva - Proceedings of the 2006 ACM symposium on Applied computing, 2006 - portal.acm.org

... that 8.5% of the doc- uments fetched were **duplicates** when using the Mercator **crawler** [23 ... due to the incremental construction of the **document** collec- tion. ...

[Cited by 12](#) - [Related articles](#) - [Web Search](#) - [All 7 versions](#)

ProFusion*: Intelligent fusion from multiple, distributed search engines- *iukm.org

S Gauch, G Wang, M Gomez - Journal of Universal Computer Science, 1996 - jucs.org

... Excite, InfoSeek, Lycos, Open Text, **WebCrawler**); ProFusion; and two ... the number of irrelevant **documents**, the number ... links, the number of **duplicates**, the number ...

[Cited by 138](#) - [Related articles](#) - [Web Search](#) - [All 20 versions](#)

[Finding near-duplicate web pages: a large-scale evaluation of algorithms- chinaunix.net \(pdf\)](#)

M Henzinger - Proceedings of the 29th annual international ACM SIGIR ..., 2006 - portal.acm.org

... it uses the same amount of space per **document** and returns ... and near- **duplicates** depends on the **crawler** used to ... to determine by hand all near-**duplicate** pairs in ...

[Cited by 50](#) - [Related articles](#) - [Web Search](#) - [All 13 versions](#)

[\[PDF\] Where and how duplicates occur in the web](#)

A Pereira Jr, R Baeza-Yates, N Ziviani - Proceedings of the Fourth Latin American Web Congress, 2006 - Citeseer

... To project a Web **crawler** many aspects must be ... the coverage and elimination of **duplicates**,

Web crawlers ... every found directory and crawling every **document** in a ...

[Cited by 1](#) - [Related articles](#) - [View as HTML](#) - [Web Search](#) - [All 5 versions](#)

[Method, system, and program for handling redirects in a search engine](#)

MF Fontoura, A Neumann, R Qi, EJ Shekita - US Patent App. 10/764,771, 2004 - Google Patents

... s) 160 s sarch Engine] [30 **Crawler** Component 132 ... Rank Component 134 Redirect Component

136 **Duplicate** Detection Component ... assigns a rank to each **document** i r ...

[Web Search](#) - [All 2 versions](#)

[Results from a web impact factor crawler- openrepository.com \(pdf\)](#)

M Theilwall - Journal of Documentation, 2001 - emeraldinsight.com

... domain. In both cases **duplicate** links in a single page to the same **document**

count only once. There ... March 2001 WIF **CRAWLER** 185 Table ...

[Cited by 46](#) - [Related articles](#) - [Web Search](#) - [BL Direct](#) - [All 8 versions](#)

Key authors: [D Hawking](#) - [N Craswell](#) - [P Bailey](#) - [S Gauch](#) - [D Fetterly](#)

Google

Result Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

duplicate document webcrawler OR

Search

[Go to Google Home](#) - [About Google](#) - [About Google Scholar](#)

©2009 Google